Market Segmentation

Jason Antal

2024-08-18

Consider the data in social_marketing.csv. This was data collected in the course of a market-research study using followers of the Twitter account of a large consumer brand that shall remain nameless—let's call it "NutrientH20" just to have a label. The goal here was for NutrientH20 to understand its social-media audience a little bit better, so that it could hone its messaging a little more sharply.

A bit of background on the data collection: the advertising firm who runs NutrientH20's online-advertising campaigns took a sample of the brand's Twitter followers. They collected every Twitter post ("tweet") by each of those followers over a seven-day period in June 2014. Every post was examined by a human annotator contracted through Amazon's Mechanical Turk service. Each tweet was categorized based on its content using a pre-specified scheme of 36 different categories, each representing a broad area of interest (e.g. politics, sports, family, etc.) Annotators were allowed to classify a post as belonging to more than one category. For example, a hypothetical post such as "I'm really excited to see grandpa go wreck shop in his geriatic soccer league this Sunday!" might be categorized as both "family" and "sports." You get the picture.

Each row of social_marketing.csv represents one user, labeled by a random (anonymous, unique) 9-digit alphanumeric code. Each column represents an interest, which are labeled along the top of the data file. The entries are the number of posts by a given user that fell into the given category. Two interests of note here are "spam" (i.e. unsolicited advertising) and "adult" (posts that are pornographic, salacious, or explicitly sexual). There are a lot of spam and pornography "bots" on Twitter; while these have been filtered out of the data set to some extent, there will certainly be some that slip through. There's also an "uncategorized" label. Annotators were told to use this sparingly, but it's there to capture posts that don't fit at all into any of the listed interest categories. (A lot of annotators may used the "chatter" category for this as well.) Keep in mind as you examine the data that you cannot expect perfect annotations of all posts. Some annotators might have simply been asleep at the wheel some, or even all, of the time! Thus there is some inevitable error and noisiness in the annotation process.

Your task to is analyze this data as you see fit, and to prepare a concise report for NutrientH20 that identifies any interesting market segments that appear to stand out in their social-media audience. You have complete freedom in deciding how to pre-process the data and how to define "market segment." (Is it a group of correlated interests? A cluster? A latent factor? Etc.) Just use the data to come up with some interesting, well-supported insights about the audience, and be clear about what you did.

```r
cat("Let's start off our analysis by simply looking at the most popular topics,
    so that NutrientH20 has a better idea of where to focus its marketing efforts.
    To accomplish this, we will remove all spam and adult tweets, then sort in
    descending order.")
```

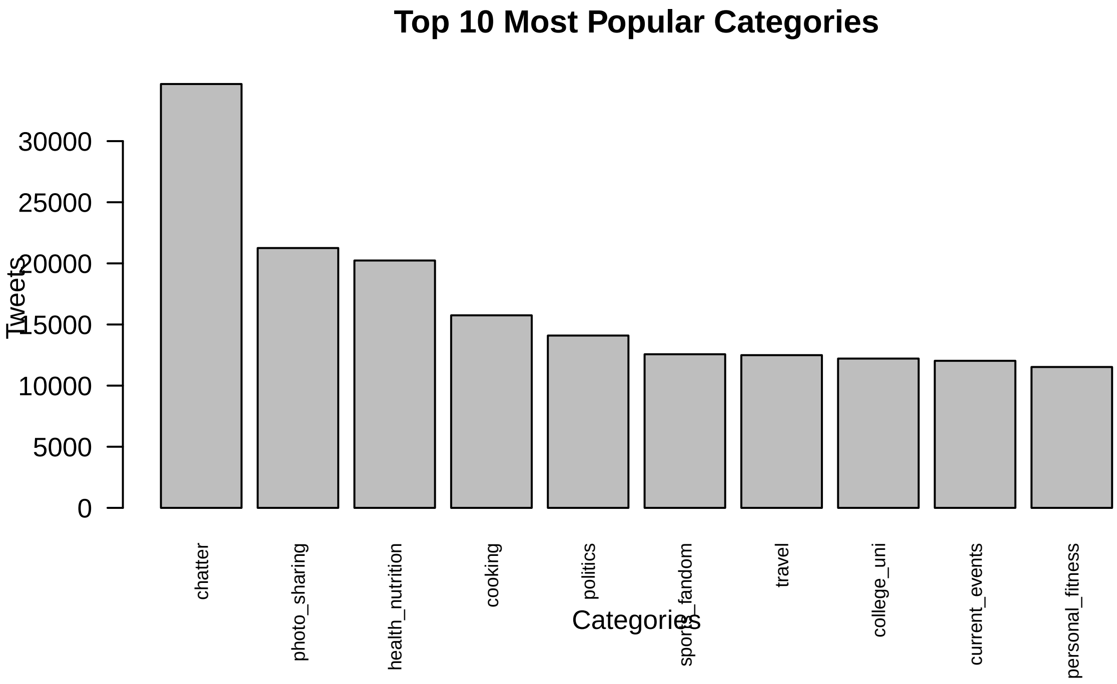
```
## Let’s start off our analysis by simply looking at the most popular topics,
##     so that NutrientH20 has a better idea of where to focus its marketing efforts.
##     To accomplish this, we will remove all spam and adult tweets, then sort in
##     descending order.
```

```
#Prepare data
names(data)[1] <- "annotator"
data$annotator <- 1:nrow(data)

data_clean <- data %>%
  select(-spam, -adult, -annotator)
data_normalized <- scale(data_clean)

#identify most popular interest categories:
category_sums <- colSums(data_clean)
top_categories <- sort(category_sums, decreasing = TRUE)[1:10]

par(mar = c(10, 5, 4, 2) + 0.1, cex = 0.8)  # Increase bottom margin for labels
barplot(top_categories,
        las = 2,
        main = "Top 10 Most Popular Categories",
        ylab = "Tweets",
        xlab = "Categories",
        cex.names = 0.7)
```



```
cat('These are the most popular topics of discussion on Twitter. From this, we
    can see that food, photo sharing, and "chatter" (which could simply mean funny
    or engaging everyday interactions) would have the widest audiences. If NutrientH20
    wants to reach its largest audience possible, it should take care not to neglect
    those topics when communicating to its audience.')
```

```
## These are the most popular topics of discussion on Twitter. From this, we
##     can see that food, photo sharing, and "chatter" (which could simply mean funny
##     or engaging everyday interactions) would have the widest audiences. If NutrientH20
##     wants to reach its largest audience possible, it should take care not to neglect
##     those topics when communicating to its audience.
```

```r
cat("It's often a good idea to focus on a few specific audiences. We will use k-means
    to create 5 clusters, then calculate the mean values of each interest category
    for each cluster. The idea is to see what the most common interests are per
    cluster, so we can better identify each market segment.")
```

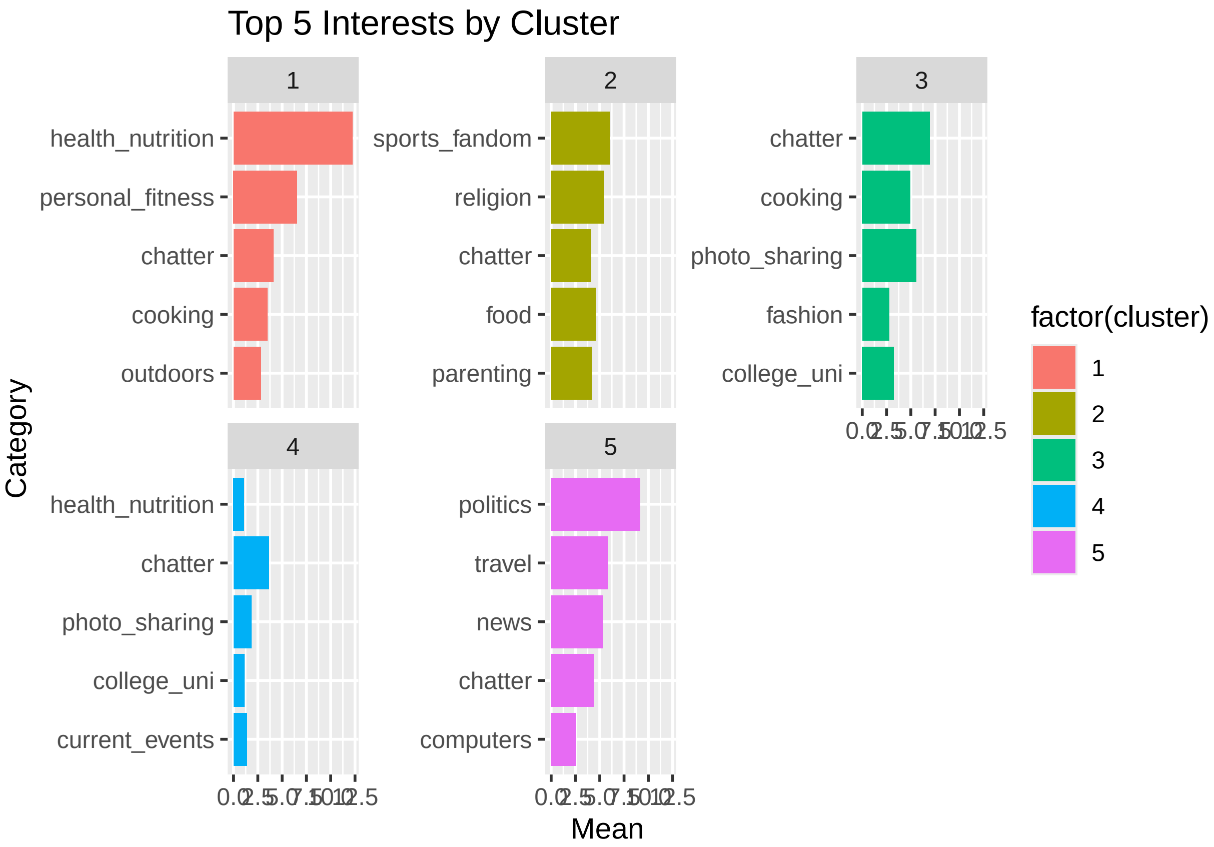
```
## It’s often a good idea to focus on a few specific audiences. We will use k-means
##     to create 5 clusters, then calculate the mean values of each interest category
##     for each cluster. The idea is to see what the most common interests are per
##     cluster, so we can better identify each market segment.
```

```r
# Perform k-means clustering to analyze data
set.seed(15)
km <- kmeans(data_normalized, centers = 5, nstart = 25)

# Calculate mean values for each cluster
cluster_means <- aggregate(data_clean, by=list(cluster=km$cluster), mean)
cluster_means_long <- gather(cluster_means, key="category", value="mean", -cluster)

# Plot the top 5 interests for each cluster
top_interests <- cluster_means_long %>%
  group_by(cluster) %>%
  top_n(5, mean) %>%
  arrange(cluster, desc(mean))

ggplot(top_interests, aes(x=reorder(category, mean), y=mean, fill=factor(cluster))) +
  geom_bar(stat="identity") +
  facet_wrap(~cluster, scales="free_y") +
  coord_flip() +
  labs(x="Category", y="Mean", title="Top 5 Interests by Cluster")
```



```r
cat("We've identified 3 distinct market segments NutrientH20 should consider targeting:
1) Health and Fitness Enthusiasts: This segment shows high interest in nutrition,
fitness, and outdoors activities. They are likely to be receptive to messages about
the health benefits of NutrientH20.
2) Global citizens: This segment is interested in travel, current events, and
politics. They might respond well to campaigns that highlight NutrientH20's
social responsibility initiatives around the world.
3) Visual engagers: This segment is interested in sharing photos, fashion,
and is also the most interested in random chatter. They would be most likely to
respond well to influencer parternships, memes, or visually appealing ad campaigns.")
```

```
## We’ve identified 3 distinct market segments NutrientH20 should consider targeting:
## 1) Health and Fitness Enthusiasts: This segment shows high interest in nutrition,
## fitness, and outdoors activities. They are likely to be receptive to messages about
## the health benefits of NutrientH20.
## 2) Global citizens: This segment is interested in travel, current events, and
## politics. They might respond well to campaigns that highlight NutrientH20’s
## social responsibility initiatives around the world.
## 3) Visual engagers: This segment is interested in sharing photos, fashion,
## and is also the most interested in random chatter. They would be most likely to
## respond well to influencer parternships, memes, or visually appealing ad campaigns.
```